US EPA TOXCAST DATA RELEASE JUNE 2022 Summary Files

This file describes the contents of the June 2022 ToxCast data release. These files are the same structure and format as the invitrodb v3.4 release, but with updates for data updates in invitrodb v3.5.

The zip file contains the following summary-level files:

```
##  [1] "ar.er.lit.Rdata"
##  [2] "ar.er.lit.xlsx"
##  [3] "ar.er.qsar.Rdata"
##  [4] "ar.er.qsar.xlsx"
##  [5] "assay_annotation_information_invitrodb_v3_5.RData"
##  [6] "assay_annotation_information_invitrodb_v3_5.xlsx"
##  [7] "assay_methods_invitrodb_v3_5.RData"
##  [8] "assay_methods_invitrodb_v3_5.xlsx"
##  [9] "Assay_Quality_Detailed_Stats.Rdata"
## [10] "Assay_Quality_Detailed_Stats_220630.csv"
## [11] "Assay_Quality_Summary_Stats.Rdata"
## [12] "Assay_Quality_Summary_Stats_220630.csv"
## [13] "Assay_Summary_220630.csv"
## [14] "CytoPt.Rdata"
## [15] "CytoPt.xlsx"
## [16] "EXPORT_LVL5&6_ASID1_ACEA_220630.csv"
## [17] "EXPORT_LVL5&6_ASID1_ACEA_220630.Rdata"
## [18] "EXPORT_LVL5&6_ASID10_VALA_220630.csv"
## [19] "EXPORT_LVL5&6_ASID10_VALA_220630.Rdata"
## [20] "EXPORT_LVL5&6_ASID11_CLD_220630.csv"
## [21] "EXPORT_LVL5&6_ASID11_CLD_220630.Rdata"
## [22] "EXPORT_LVL5&6_ASID12_CCTE_PADILLA_220630.csv"
## [23] "EXPORT_LVL5&6_ASID12_CCTE_PADILLA_220630.Rdata"
## [24] "EXPORT_LVL5&6_ASID13_TANGUAY_220630.csv"
## [25] "EXPORT_LVL5&6_ASID13_TANGUAY_220630.Rdata"
## [26] "EXPORT_LVL5&6_ASID14_STM_220630.csv"
## [27] "EXPORT_LVL5&6_ASID14_STM_220630.Rdata"
## [28] "EXPORT_LVL5&6_ASID16_ARUNA_220630.csv"
## [29] "EXPORT_LVL5&6_ASID16_ARUNA_220630.Rdata"
## [30] "EXPORT_LVL5&6_ASID17_CCTE_220630.csv"
## [31] "EXPORT_LVL5&6_ASID17_CCTE_220630.Rdata"
## [32] "EXPORT_LVL5&6_ASID2_APR_220630.csv"
## [33] "EXPORT_LVL5&6_ASID2_APR_220630.Rdata"
## [34] "EXPORT_LVL5&6_ASID20_CCTE_SHAFER_220630.csv"
## [35] "EXPORT_LVL5&6_ASID20_CCTE_SHAFER_220630.Rdata"
## [36] "EXPORT_LVL5&6_ASID21_CPHEA_STOKER_220630.csv"
## [37] "EXPORT_LVL5&6_ASID21_CPHEA_STOKER_220630.Rdata"
## [38] "EXPORT_LVL5&6_ASID24_CCTE_GLTED_220630.csv"
## [39] "EXPORT_LVL5&6_ASID24_CCTE_GLTED_220630.Rdata"
## [40] "EXPORT_LVL5&6_ASID25_UPITT_220630.csv"
## [41] "EXPORT_LVL5&6_ASID25_UPITT_220630.Rdata"
## [42] "EXPORT_LVL5&6_ASID27_UKN_220630.csv"
## [43] "EXPORT_LVL5&6_ASID27_UKN_220630.Rdata"
## [44] "EXPORT_LVL5&6_ASID28_ERF_220630.csv"
## [45] "EXPORT_LVL5&6_ASID28_ERF_220630.Rdata"
## [46] "EXPORT_LVL5&6_ASID29_TAMU_220630.csv"
## [47] "EXPORT_LVL5&6_ASID29_TAMU_220630.Rdata"
```

```
## [48] "EXPORT_LVL5&6_ASID3_ATG_220630.csv"
## [49] "EXPORT_LVL5&6_ASID3_ATG_220630.Rdata"
## [50] "EXPORT_LVL5&6_ASID30_IUF_220630.csv"
## [51] "EXPORT_LVL5&6_ASID30_IUF_220630.Rdata"
## [52] "EXPORT_LVL5&6_ASID31_CCTE_MUNDY_220630.csv"
## [53] "EXPORT_LVL5&6_ASID31_CCTE_MUNDY_220630.Rdata"
## [54] "EXPORT_LVL5&6_ASID4_BSK_220630.csv"
## [55] "EXPORT_LVL5&6_ASID4_BSK_220630.Rdata"
## [56] "EXPORT_LVL5&6_ASID5_NVS_220630.csv"
## [57] "EXPORT_LVL5&6_ASID5_NVS_220630.Rdata"
## [58] "EXPORT_LVL5&6_ASID6_OT_220630.csv"
## [59] "EXPORT_LVL5&6_ASID6_OT_220630.Rdata"
## [60] "EXPORT_LVL5&6_ASID7_TOX21_220630.csv"
## [61] "EXPORT_LVL5&6_ASID7_TOX21_220630.Rdata"
## [62] "EXPORT_LVL5&6_ASID8_CEETOX_220630.csv"
## [63] "EXPORT_LVL5&6_ASID8_CEETOX_220630.Rdata"
## [64] "EXPORT_LVL5&6_ASID9_LTEA_220630.csv"
## [65] "EXPORT_LVL5&6_ASID9_LTEA_220630.Rdata"
## [66] "gene_target_information_invitrodb_v3_5.RData"
## [67] "gene_target_information_invitrodb_v3_5.xlsx"
## [68] "ht.h295r.model.Rdata"
## [69] "ht.h295r.model.xlsx"
## [70] "sc1_sc2_invitrodb_v3_5.Rdata"
## [71] "sc1_sc2_invitrodb_v3_5.xlsx"
## [72] "toxcast_ar_pathway_model_scores.Rdata"
## [73] "toxcast_ar_pathway_model_scores.xlsx"
## [74] "toxcast_er_pathway_model_scores.Rdata"
## [75] "toxcast_er_pathway_model_scores.xlsx"
```

In addition to the above listed files, the ToxCast program also released a MySQL dump file containing all data and a beta version of the R package (tcpl) that interacts with the MySQL database used to process all of the data for this release. For information/data not included in the listed summary files, users will need to download and interact with the MySQL database. We also encourage the database users to utilize the 'tcpl' R package containing numerous queries and functionality for easily loading and visualizing the data. For more information on how to on data retrieval with tcpl please see the data retrieval vignette. https://cran.r-project.org/web/packages/tcpl/vignettes/Data_retrieval.html

Each section below will describe a subset of the summary level information provided. Each of these files are provided in xlsx/csv as well as Rdata to allow users to more easily interact with the data using the R programming language.

Assay Information

```
## [1] "assay_annotation_information_invitrodb_v3_5.RData"
## [2] "assay_annotation_information_invitrodb_v3_5.xlsx"
## [3] "assay_methods_invitrodb_v3_5.RData"
## [4] "assay_methods_invitrodb_v3_5.xlsx"
## [5] "Assay_Quality_Detailed_Stats.Rdata"
## [6] "Assay_Quality_Detailed_Stats_220630.csv"
## [7] "Assay_Quality_Summary_Stats.Rdata"
## [8] "Assay_Quality_Summary_Stats_220630.csv"
## [9] "Assay_Summary_220630.csv"
```

The assay annotation information file contains all of the annotation fields used to describe an assay, assay component, and assay component endpoint, including assay citation, assay reagent, and assay target information.

The definition of an "assay" is, for the purposes of this package, broken into:

- assay_source – the vendor/origination of the data
- assay – the procedure to generate the component data
- assay_component – the raw data readout(s)
- assay_component_endpoint – the normalized component data

Each assay element is represented by a separate table in the tcpl database. In general, we refer to an "assay_component_endpoint" as an "assay endpoint." As we move down the hierarchy, each additional layer has a one-to-many relationship with the previous layer. For example, an assay component can have multiple assay endpoints, but an assay endpoint can derive only from a single assay component.

All processing occurs by assay component or assay endpoint, depending on the processing type (single-concentration or multiple-concentration) and level. No data are stored at the assay or assay source level. The "assay" and "assay_source" tables store annotations to help in the processing and down-stream understanding/analysis of the data.

Source: https://cran.r-project.org/web/packages/tcpl/vignettes/Introduction_Appendices.html

The assay methods invitrodb file contains all of the methods applied to each of the assays during each step of the pipeline.

The assay endpoint detailed statistics are derived from the raw concentration response data and provide assay-plate-wise statistics common to the high throughput screening community, including z-prime and ssmd (strictly standardized mean difference). The detailed file provides the median and median absolute deviation across all plates, where applicable. These calculations are performed at the assay component level (i.e., assay readout) rather than the endpoint (i.e., direction of analysis).

- acid = assay component id (unique id for each assay readout)
- acnm = assay component name
- nmed = neutral control well median value, by plate
- nmad = neutral control median absolute deviation, by plate
- pmed = positive control well median value, by plate
- pmad = positive control well median absolute deviation, by plate
- mmed = negative control well median, by plate
- mmad = negative control well median absolute deviation value, by plate
- zprm.p = robust z-prime, median across all plates using positive control wells
- zprm.m = robust z-prime, median across all plates using negative control wells
- ssmd.p = robust ssmd, median across all plates using positive control wells
- ssmd.m = robust ssmd, median across all plates using negative control wells
- cv = median coefficient of variation across all plate
- sn.p = median signal-to-noise across all plates based on positive control wells
- sn.m = median signal-to-noise across all plates based on negative control wells
- sb.p = median signal-to-background across all plates based on positive control wells
- sb.m = median signal-to-background across all plates based on negative control wells

Many of these calculations result in NA values either because plate-level details were not provided or because the analysis precludes this calculation. This initial release of the quality statistics are for general and relative reference only. Due to the diverse assay technologies and study designs deployed, a highly generalized and robust (median and mad vs mean and sd) set of calculations were performed.

- aeid = assay endpoint id (unique id)
- ocnc = overall concordance among chemical replicates calculated as the percentage of time all samples for a chemical were either negative or positive (e.g., 0 out of 3 or 3 out of 3) over the total number of chemicals with replicates.
- hcnc = hit concordance among chemical replicates calculated as the percentage of time all samples for a chemical were positive (e.g., 3 out of 3) over the total number of chemicals with any replicate being positive (e.g., 1 out of 3 or 2 out of 3). *It should be noted that most of these chemical replicates were separately procured and that these concordance values are highly influenced by the number of replicates.*
- aenm = assay endpoint name (i.e., assay_component_endpoint_name)
- resp_unit = response unit (fold induction or percent activity)
- bmad = baseline median absolute deviation for the assay (based on the response values at the 2 lowest tested concentrations)
- nconc = nominal number of tested concentrations
- coff = the response cutoff used to derive the hit calls (e.g., 5*bmad, 10*bmad)
- test = total number of samples tested
- acnt = number of active samples
- apct = percent active samples
- icnt = number of inactive samples
- ipct = percent of inactive samples
- ncnt = number of samples that could not be modeled (e.g., having less than 4 concs)
- npct = percent not modeled
- mmed = maximum observed response across the assay
- cmax = target (nominal) maximal tested concentration
- cmin = target (nominal) minimal tested concentration
- mtop = maximum modeled response across the assay (max top of curve)
- nrep = target (nominal) number of replicates
- npts = target (nominal) number of points (nconc //* nrep)
- cnst = percent constant model winner (based on having lowest AIC value)
- hill = percent hill model winner (based on having lowest AIC value)
- gnls = percent gain-loss model winner (based on having lowest AIC value)
- rmse = median root mean squared error across all winning models

The summary quality statistics file provides a nice overview of the target study design for each assay endpoint as well as summary statistics around active prevalence and hit-calling criteria.

Gene coverage (gene + intended_target)

```
## [1] "gene_target_information_invitrodb_v3_5.RData"
## [2] "gene_target_information_invitrodb_v3_5.xlsx"
```

The above gene target information file includes all of the annotated information about particular genes that each assay component endpoint targets.

Single concentration summary (sc1 + sc2)

```
## [1] "sc1_sc2_invitrodb_v3_5.Rdata" "sc1_sc2_invitrodb_v3_5.xlsx"
```

These files provide the level 1 and level 2 information for all chemicals tested at a single concentration. Generally, the goal of single-concentration processing is to identify potentially active compounds from a broad screen at a single concentration.

Level 1 processing converts the assay component to assay endpoint(s) and defines the normalized-response value field (resp), logarithm-concentration field (logc), and optionally, the baseline value (bval) and positive

control value (pval) fields. The purpose of level 1 is to normalize the raw values to either the percentage of a control or to fold-change from baseline.

Level 2 processing defines the baseline median absolute deviation (bmad), collapses any replicates by sample ID, and determines the activity.

Before the data are collapsed by sample ID, the bmad is calculated as the median absolute deviation of all wells with well type equal to "t." The calculation to define bmad is done once across the entire assay endpoint. If additional data is added to the database for an assay component, the bmad values for all associated assay endpoints will change. Note, this bmad definition is different from the bmad definition used for multiple-concentration screening.

To collapse the data by sample ID, the median response value is calculated at each concentration. The data are then further collapsed by taking the maximum of those median values (max_med).

Once the data are collapsed, such that each assay endpoint-sample pair only has one value, the activity is determined. For a sample to get an active hit call, the max_med must be greater than an efficacy cutoff. The efficacy cutoff is determined by the level 2 methods. The efficacy cutoff value (coff) is defined as the maximum of all values given by the assigned level 2 methods. Failing to assign a level 2 method will result in every sample being called active.

Below is a list of important columns and a breif definition: s1id - Level 1 ID s0id - Level 0 ID acid - Assay component ID aeid - Assay component endpoint ID logc - Log base 10 concentration bval - Baseline value pval - Positive control value resp - Normalized response value

s2id - Level 2 ID aeid - Assay component endpoint ID spid - Sample ID bmad - Baseline median absolute deviation max_med - Maximum median response value hitc - Hit-/activity-call, 1 if active, 0 if inactive coff - Efficacy cutoff value tmpi - Ignore, temporary index used for uploading purposes

Multi-concentration Curve-fitting Information (Levels 5,6,7)

```
##  [1] "EXPORT_LVL5&6_ASID1_ACEA_220630.csv"
##  [2] "EXPORT_LVL5&6_ASID1_ACEA_220630.Rdata"
##  [3] "EXPORT_LVL5&6_ASID10_VALA_220630.csv"
##  [4] "EXPORT_LVL5&6_ASID10_VALA_220630.Rdata"
##  [5] "EXPORT_LVL5&6_ASID11_CLD_220630.csv"
##  [6] "EXPORT_LVL5&6_ASID11_CLD_220630.Rdata"
##  [7] "EXPORT_LVL5&6_ASID12_CCTE_PADILLA_220630.csv"
##  [8] "EXPORT_LVL5&6_ASID12_CCTE_PADILLA_220630.Rdata"
##  [9] "EXPORT_LVL5&6_ASID13_TANGUAY_220630.csv"
## [10] "EXPORT_LVL5&6_ASID13_TANGUAY_220630.Rdata"
## [11] "EXPORT_LVL5&6_ASID14_STM_220630.csv"
## [12] "EXPORT_LVL5&6_ASID14_STM_220630.Rdata"
## [13] "EXPORT_LVL5&6_ASID16_ARUNA_220630.csv"
## [14] "EXPORT_LVL5&6_ASID16_ARUNA_220630.Rdata"
## [15] "EXPORT_LVL5&6_ASID17_CCTE_220630.csv"
## [16] "EXPORT_LVL5&6_ASID17_CCTE_220630.Rdata"
## [17] "EXPORT_LVL5&6_ASID2_APR_220630.csv"
## [18] "EXPORT_LVL5&6_ASID2_APR_220630.Rdata"
## [19] "EXPORT_LVL5&6_ASID20_CCTE_SHAFER_220630.csv"
## [20] "EXPORT_LVL5&6_ASID20_CCTE_SHAFER_220630.Rdata"
## [21] "EXPORT_LVL5&6_ASID21_CPHEA_STOKER_220630.csv"
## [22] "EXPORT_LVL5&6_ASID21_CPHEA_STOKER_220630.Rdata"
## [23] "EXPORT_LVL5&6_ASID24_CCTE_GLTED_220630.csv"
## [24] "EXPORT_LVL5&6_ASID24_CCTE_GLTED_220630.Rdata"
## [25] "EXPORT_LVL5&6_ASID25_UPITT_220630.csv"
```

```
## [26] "EXPORT_LVL5&6_ASID25_UPITT_220630.Rdata"
## [27] "EXPORT_LVL5&6_ASID27_UKN_220630.csv"
## [28] "EXPORT_LVL5&6_ASID27_UKN_220630.Rdata"
## [29] "EXPORT_LVL5&6_ASID28_ERF_220630.csv"
## [30] "EXPORT_LVL5&6_ASID28_ERF_220630.Rdata"
## [31] "EXPORT_LVL5&6_ASID29_TAMU_220630.csv"
## [32] "EXPORT_LVL5&6_ASID29_TAMU_220630.Rdata"
## [33] "EXPORT_LVL5&6_ASID3_ATG_220630.csv"
## [34] "EXPORT_LVL5&6_ASID3_ATG_220630.Rdata"
## [35] "EXPORT_LVL5&6_ASID30_IUF_220630.csv"
## [36] "EXPORT_LVL5&6_ASID30_IUF_220630.Rdata"
## [37] "EXPORT_LVL5&6_ASID31_CCTE_MUNDY_220630.csv"
## [38] "EXPORT_LVL5&6_ASID31_CCTE_MUNDY_220630.Rdata"
## [39] "EXPORT_LVL5&6_ASID4_BSK_220630.csv"
## [40] "EXPORT_LVL5&6_ASID4_BSK_220630.Rdata"
## [41] "EXPORT_LVL5&6_ASID5_NVS_220630.csv"
## [42] "EXPORT_LVL5&6_ASID5_NVS_220630.Rdata"
## [43] "EXPORT_LVL5&6_ASID6_OT_220630.csv"
## [44] "EXPORT_LVL5&6_ASID6_OT_220630.Rdata"
## [45] "EXPORT_LVL5&6_ASID7_TOX21_220630.csv"
## [46] "EXPORT_LVL5&6_ASID7_TOX21_220630.Rdata"
## [47] "EXPORT_LVL5&6_ASID8_CEETOX_220630.csv"
## [48] "EXPORT_LVL5&6_ASID8_CEETOX_220630.Rdata"
## [49] "EXPORT_LVL5&6_ASID9_LTEA_220630.csv"
## [50] "EXPORT_LVL5&6_ASID9_LTEA_220630.Rdata"
```

The above files summarise levels 5,6,7 of the multi-concentration to estimate the activity, potency, efficacy, and other parameters for sample-assay pairs.

Level 5 processing determines the winning model and activity for the concentration series, bins all of the concentration series into categories, and calculates additional point-of-departure estimates based on the activity cutoff. A level 5 method must be assigned for hit-calling, as at least one median response from a tested concentration must exceed the cutoff for a positive hit-call.

Level 6 processing uses various methods to identify concentration series with etiologies that may suggest false positive/false negative results or explain apparent anomalies in the data. Each flag is defined by a level 6 method that has to be assigned to each assay endpoint. An assay endpoint does not need any level 6 methods assigned to complete processing.

Level 7 processing implements smooth nonparametric bootstrapping, a statistical method that uses resampling and added noise to determine uncertainty in a series. Due to the binary, and semi-arbitrary nature of activity cutoffs, it may be hard to determine by hit-call alone the confidence of a curve fit. By adding random normally distributed noise to the series, if similar results are produced, one could be more confident in the results. By resampling and adding normally distributed noise over many iterations, a general picture of the confidence in a curve fit can be ascertained. This is all generated using the toxboot R package: https://github.com/ericwatt/toxboot

All information in the summary file is reported at the sample level and a single file has been produced per assay source name (asnm) or assay source id (asid). Each row in this file contains a unique combination of sample (spid) and assay endpoint (aeid) with all of the model information applied to the underlying concentration response data.

Additional information about data processing can be found in the tcpl vignettes:

Introduction - https://cran.r-project.org/web/packages/tcpl/vignettes/Introduction_Appendices.html

Data Processing - https://cran.r-project.org/web/packages/tcpl/vignettes/Data_processing.html

All columns starting with 'cnst' refer to parameters from the constant model. All columns starting with 'hill' refer to parameters from the hill model. All columns starting with 'gnls' refer to parameters from the gain-loss model (the product of 2 hill models, one with a negative hill slope, and that share a top/upper-asymptote). For more information on the specifics of the modeling process please refer to the data analysis R package documentation (Filer et al. 2017 DOI: https://doi.org/10.1093/bioinformatics/btw680). All columns starting with 'modl' refer to the winning models parameters. Below is the list of the 91 columns exported for this dataset from Level 4 (modeling), 5 (model selection & hit calling), 6 (flagging) data processing and 7 (statistical bootstrapping).

1. m5id = unique id level 5 processing
2. spid = sample id (id blindly provided to vendors)
3. chid = chemical id (DSSTox GSID) 1:1 with casrn
4. casn = CAS Registry number
5. chnm = chemical name
6. code = CAS Registry number (excel protected)
7. aeid = assay endpoint id (unique id)
8. aenm = assay endpoint name
9. m4id = unique id for level 4 processing
10. bmad = baseline median absolute deviation (noise around baseline)
11. resp_max = maximal single replicate response
12. resp_min = minimal single replicate response
13. max_mean = maximal mean response at a given concentration
14. max_mean_conc = corresponding concentration of max_mean
15. max_med = maximal median response at a given concentration
16. max_med_conc = corresponding concentration of max_med
17. logc_max = maximum tested log concentration (log uM)
18. logc_min = minimum tested log concentration (log uM)
19. cnst = constant model successfully run (1 or 0)
20. hill = hill model successfully run (1 or 0)
21. hcov = hill model covariance
22. gnls = gain-loss model successfully run (1 or 0)
23. gcov = gain-loss model covariance
24. cnst_er = constant model error term
25. cnst_aic = constant model AIC (used to select winning model)
26. cnst_rmse = constant model RMSE
27. cnst_prob = constant model probability (based on AIC)
28. hill_tp = hill model top of curve
29. hill_tp_sd = hill model top standard deviation
30. hill_ga = hill model logAC50 (gain logAC50)
31. hill_ga_sd = hill model AC50 standard deviation
32. hill_gw = hill model slope
33. hill_gw_sd = hill model slope standard deviation
34. hill_er = hill model error term
35. hill_er_sd = hill model error standard deviation
36. hill_aic = hill model AIC (used to select winning model)
37. hill_rmse = hill model RMSE
38. hill_prob = hill model probability (based on AIC)
39. gnls_tp = gain-loss top of curve
40. gnls_tp_sd = gain-loss top of curve standard deviation
41. gnls_ga = gain-loss model gain logAC50
42. gnls_ga_sd = gain-loss model gain logAC50 standard deviation
43. gnls_gw = gain-loss model gain slope (positive)
44. gnls_gw_sd = gain-loss model gain slope standard deviation
45. gnls_la = gain-loss model loss logAC50

46. gnls_la_sd = gain-loss model loss logAC50 standard deviation
47. gnls_lw = gain-loss model loss slope (negative)
48. gnls_lw_sd = gain-loss model loss slope standard deviation
49. gnls_er = gain-loss model error term
50. gnls_er_sd = gain-loss model error standard deviation
51. gnls_aic = gain-loss model AIC (used to select model winner)
52. gnls_rmse = gain-loss model RMSE
53. gnls_prob = gain-loss model probability (based on AIC)
54. nconc = number of tested concentrations
55. npts = number of data points
56. nrep = number of replicates
57. nmed_gtbl = number of median values greater than baseline
58. hitc = hit call (based on 'coff' and winning model) positive =1, negative =0, not enough data to fit = -1; max_med must exceed coff
59. modl = winning model
60. fitc = fit category (defined by many parameters)
61. coff = response cutoff (used to define hit-call)
62. actp = activity probability (1-cnst_prob)
63. modl_er = winning model error term
64. modl_tp = winning model top of curve (where applicable)
65. modl_ga = winning model gain logAC50 (where applicable)
66. modl_gw = winning model gain slope (where applicable)
67. modl_la = winning model loss logAC50 (where applicable)
68. modl_lw = winning model loss slope (where applicable)
69. modl_rmse = winning model RMSE
70. modl_prob = winning model probability
71. modl_acc = winning model log concentration at 'coff'
72. modl_acb = winning model log concentration at 'bmad'
73. resp_unit = response units
74. flag_id = concatenated list of flag ids
75. flag = concatenated list of flag names
76. chit = chemical-level hit call
77. stkc = stock concentration of sample
78. stkc_unit = stock concentration unit, typically mM
79. test_conc_unit = tested concentration unit, typically uM
80. spid_legacy = legacy sample id
81. gsid_rep = representative sample; based on tcplSubsetChid()
82. hit_pct = Total percent of hit calls made after 1000 bootstraps
83. total_hitc = Total number of hit calls made after 1000 bootstraps
84. modl_ga_min = Low bound of the 95% confidence interval for the AC50
85. modl_ga_max = Upper bound of the 95% confidence interval for the AC50
86. modl_ga_med = Median AC50 after 1000 bootstraps
87. modl_gw_med = Median gain Hill coefficient for 1000 bootstraps
88. modl_ga_delta = AC50 confidence interval width in log units
89. cnst_pct = Percent of 1000 bootstraps that the constant model was selected as the winning model
90. hill_pct = Percent of 1000 bootstraps that the Hill model was selected as the winning model
91. gnls_pct = Percent of 1000 bootstraps that the gain-loss was selected as the winning model

The parameters for the winning model are given regardless of hit-calling (also known as an activity call); therefore, many inactive chemicals have a gain AC50 chemical in the "modl_ga" column, for example. The "hitc" column provides the activity call (1=active, 0=inactive, -1=unable to model)

Model information

```
##  [1] "ar.er.lit.Rdata"
##  [2] "ar.er.lit.xlsx"
##  [3] "ar.er.qsar.Rdata"
##  [4] "ar.er.qsar.xlsx"
##  [5] "ht.h295r.model.Rdata"
##  [6] "ht.h295r.model.xlsx"
##  [7] "toxcast_ar_pathway_model_scores.Rdata"
##  [8] "toxcast_ar_pathway_model_scores.xlsx"
##  [9] "toxcast_er_pathway_model_scores.Rdata"
## [10] "toxcast_er_pathway_model_scores.xlsx"
```

The above files detail the results of running various models based on data available in the toxcast dataset.

- Ht.h295r.model.RData – HT-H295R Mahalanobis distance model information, as published in Haggard et al. 2018, DOI: 10.1093/toxsci/kfx274
- Ar.er.li.xlsx/Rdata: COMPARA and CERAPP literature information (as published in 10.1289/EHP5580 and 10.1289/ehp.1510267)
  - Dtxsid, casrn, name, canonical SMILES, inChi key: chemical substance identifiers
  - Literature_mode: ER or AR, and agonist, antagonist, or binding
  - Literature_score: qualitative descriptors of the literature evidence for interaction, NA, Inactive, Very Weak, Weak, Medium, Moderate, Strong
- Ar.er.qsar.xlsx/Rdata: CERAPP and COMPARA consensus QSAR model scores (as published in 10.1289/EHP5580 and 10.1289/ehp.1510267)
  - Dtxsid, casrn, name, canonical SMILES, inChi key: chemical substance identifiers
  - Qsar_mode: ER or AR, and agonist, antagonist, or binding
  - Qsar_score: ER or AR, and agonist, antagonist, or binding
- ToxCast ER and AR Pathway model scores (as published in 10.1093/toxsci/kfv168 and 10.1021/acs.chemrestox.6b00347)
  - dtxsid, casrn, name: chemical substance identifiers
  - auc_agonist: area under the curve score for agonist mode; >0.1 = positive; 0.001-0.1 = equivocal; negatives < 0.001
  - auc_antagonist: area under the curve score for antagonist mode; >0.1 = positive; 0.001-0.1 = equivocal; negatives < 0.001

Cytotoxicity Table

```
## [1] "CytoPt.Rdata" "CytoPt.xlsx"
```

CytoPt calculates the cytotoxicity point and average cytotoxicity distribution based on the activity in the "burst" assay endpoints.

- "chid" – The chemical ID; this is an index specific to invitrodb.
- "code" – The chemical code; this is an unformatted casn, specific to invitrodb.
- "chnm" – The chemical name.
- "casn" – The chemical CASRN.
- "med" – The median of the "burst" endpoint log(AC50) ("modl_ga" in the level 5 output) values.
- "mad" – The MAD of the "burst" endpoint log(AC50) values.
- "ntst" – The number of "burst" endpoints tested.
- "nhit" – The number of active "burst" endpoints.

- "use_global_mad" – TRUE/FALSE, whether the mad value was used in the global MAD calculation.
- "global_mad" – The median of the "mad" values where "use_global_mad" is TRUE.
- "cyto_pt" – The cytotoxicity point, or the value in "med" when "nhit" is at least 5% of the total assay endpoints
- "cyto_pt_um" – 10^cyto_pt
- "lower_bnd_um" – 10^(cyto_pt - 3*global_mad) (micromolar units as displayed on CompTox Chemicals Dashboard)

This table can be joined with the sample or chemical tables to connect the chemical identity with a sample identifier and/or DSSTox identifier.

For questions or concerns, please contact Jason Brown at: brown.jason@epa.gov.